Sequencing Gloss

Evolving Terminology for Emerging Technol
Comments? Suggestions Revisions? mchitty@healthtech
Last revised December 28, :

View a Printer-Friendly Version of this Web Page!

Email this to a Colle

Yes No

Fax:

SUBMIT RESET

The "race" to sequence the Human Genome is not a 100 yard dash, but a marathon. Although the ***Human Genome Project*** *is v ahead of schedule, and a number of genes have been identified, we have just begun to get a glimpse of what specific genes do a how we might be able to better use this knowledge for therapeutic interventions. Teasing apart the interactions of genes and proteins, delineating changes throughout the* ***cell cycle****, and correlating changes with health and disease will take even more tim But with complete sequences, and the possibility of cross- species comparisons we can expect new insights and speeding up ov time. And sequencing DNA is only a first step towards finding what functions are connected with specific sequences. Sequencing proteins (and then determining the structure of proteins – and the function of proteins) comes next. Progress is being made with proteins, but sequencing of carbohydrates is even more difficult.*

Applications Technologies Map: Finding guide to terms in these glossaries Site Map
Related glossaries include Applications Functional genomics, Proteomics
Informatics Algorithms, Bioinformatics, In silico & molecular modeling
Technologies Chromatography & electrophoresis, Mass spectrometry
Biology Proteins, Protein Structures, SNPs & genetic variations, Sequences - DNA & beyond

alignment: The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology. [NCBI Bioinformatic

)arrower terms: **global alignment, local alignment, optimal alignment, pairwise alignment. Related terms BLAST, BEAU BLAST2, FASTA, gapped BLAST, Needleman - Wunsch, Smith - Waterman alignment**

annotation: Bioinformatics glossary

assembled: The term used to describe the process of using a computer to join up bits of sequence into a larger whole. [Peer Bor Richard Copley "Filling in the gaps" Nature 409: 218-820, 15 Feb. 2001]

This is different from assembly language, and the source of some confusion between biologists and computer scientists.

Related term contig assembly

automation: See sequencers- automation

BEAUTY BLAST Enhanced Alignment Utility: An enhanced version of the NCBI's BLAST database search tool. BEAUTY, wh used to search three new custom sequence databases that we have developed, incorporates information on sequence family membership, the location of the conserved domains, and the locations of any annotated domains and sites directly into BLAST se results. These enhancements make it much easier to detect weak, but functionally significant, matches in BLAST database searc http://searchlauncher.bcm.tmc.edu:9331/seq-search/Help/beauty.html

BLAST (Basic Local Alignment Search Tool): Software program from NCBI for searching public databases for homologous sequences or proteins. Designed to explore all available sequence databases regardless of whether query is protein or DNA. http://www.ncbi.nlm.nih.gov/BLAST/

aster but less rigorous than **FASTA or Smith- Waterman**

BLAST2: A newer release of BLAST that allows insertions or deletions in the aligned sequences. Gapped alignments may be mc biologically significant. Synonymous with **gapped BLAST**

BLAT: BLAST Like Alignment Tool, Jim Kent http://www.genomeblat.com/genomeblat/index.asp

base caller: A programs that analyzes trace data in chromatogram files and assigns a base for each peak. Some programs, for example, Phred and TraceTuner also produce a corresponding quality value for each base. [DNA Sequence Glossary Geospiza] http://www.geospiza.com/community/support/glossary/.

cDNA sequencing: Within the DOE [US Dept. of Energy] Human Genome Program, support began in 1990 for high- throughput analyses of **cDNAs**, the sturdy representatives of the cell's fragile **messenger RNAs (mRNAs)** for gene expression. In 1994 I.M.A.G.E was founded to coordinate worldwide, public- sector efforts in cDNA analyses. For a few years, most cDNA sequencin used to catalogue cDNAs by short sequence reads called **expressed sequence tags (ESTs)**. In the fall of 1996, several researc teams worldwide announced plans for complete sequencing of cDNAs. The following spring, a series of international cDNA work began. Building on the **Integrated Molecular Analysis of Genome Expression (I.M.A.G.E.)** consortium, the workshops' objecti are to extend the infrastructure I.M.A.G.E. has provided since 1994 to the challenges of complete cDNA sequencing. http://www.oml.gov/meetings/wccs/index.html

chain termination sequencing method: See Sanger sequencing (under Maxam- Gilbert & Sanger).

chemical cleavage sequencing: See Maxim- Gilbert sequencing.

chemical degradation sequencing: See Maxim- Gilbert sequencing

chromosome-specific shotgun sequencing: Chromosomes are separated by pulse- field gel electrophoresis and purified DNA used to constructed small- insert shotgun libraries. Clones from these libraries are subsequently sequenced. *Plasmodium falcipa* Genome Project Sequencing Strategies, NCBI, US, 2001 http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/pfalciparum1.html

clone, cloning: Cell biology glossary Related term: library

clone pickers: A number of institutions and biotech companies have designed robotic "clone pickers" to automate this picking an sorting process. The collection of clones is grown on large media plates, and optical scanners detect individual clones for the pick arm. The robotic arm picks each clone with a needle, transfers the clone to growth media arrayed in 96- or 384- well **microtiter plates**, sterilizes the needle and continues the picking process. Commercial clone pickers capable of picking and sorting over 10(clones per hour into arrayed microtiter plates are available. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001.

consensus sequence: A theoretical representative nucleotide or amino acid sequence in which each nucleotide or amino acid is one, which occurs most frequently at that site in the different forms which occur in nature. The phrase also refers to an actual sequence, which approximates the theoretical consensus. A known CONSERVED SEQUENCE set is represented by a consensu sequence. Commonly observed supersecondary protein structures (AMINO ACID MOTIFS) are often formed by **conserved sequences**. [MeSH, 1991]

A sequence of DNA, RNA, protein or carbohydrate derived from a number of similar molecules, which comprises the essential features for a particular function. [IUPAC Bioinorganic]

Related term: Protein structures: amino acid motifs

conserved sequence: A sequence of amino acids in a polypeptide or of nucleotides in DNA or RNA that is similar across multipl species. A known set of conserved sequences is represented by a **CONSENSUS SEQUENCE**. AMINO ACID MOTIFS are often composed of conserved sequences. [MeSH, 1993]

A "highly conserved sequence" is a DNA sequence that is very similar in several different kinds of organisms. Scientists regard th cross species similarities as evidence that a specific gene performs some basic **function** essential to many forms of life and that

evolution has therefore conserved its structure by permitting few **mutations** to accumulate in it. [NHGRI]

contig: Group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. [DOE]

A contiguous (i.e. without gaps) stretch of DNA sequence which has been assembled solely on the basis of direct sequencing information, i.e. sequencer reads. Note however that 'contig' is used in other contexts in genomics to mean a contiguous assemb something (e.g. clones), without necessarily implying that all the bases in the assembly have been determined. Ensembl Glossar http://vega.sanger.ac.uk/Mus_musculus/helpview?se=1&kw=glossary

Narrower terms: initial sequence contigs, merged sequence contigs. Related terms clone, contig assembly, scaffolds.

Published genome sequence has many gaps and interruptions. Concept of "contig" is crucial to our understanding of current limitations. [David Galas "Making sense of the sequence" Science 291 (5507): 1257, Feb. 16, 2001]

contig assembly: One of the most difficult and critical functions in DNA sequence analysis is putting together fragments from sel overlapping segments. Some programs do this better than others, particularly when dealing with sequences containing gaps. [La De Francesco "Some things considered" Scientist 12[20]:18, Oct. 12, 1999] http://www.the-scientist.com/yr1998/oct/profile1_981012.html

NCBI Contig Assembly and Annotation Process, National Center for Biotechnology Information, US, , Feb. 2001 http://www.ncbi.nlm.nih.gov/genome/guide/build.html

contig mapping: **Maps glossary**

coverage (or depth): The average number of times that a nucleotide is represented by a high- quality base in a collection of rand raw sequence. Operationally, a `high- quality base' is defined as one with an accuracy of at least 99% (corresponding to a Phred score of at least 20). [UC-Santa Cruz, US, Human Genome Project Working Draft Terminology, 2001] http://genome.ucsc.edu/goldenPath/term.html

DNA library: **Combinatorial libraries & synthesis**

DNA reaction setup: Each individual DNA sequencing project defines the criteria for reaction setup, which include: 1) the type of template; 2) the amount of DNA sequence that needs to be determined from each template; and 3) how many templates are bein analyzed. Quantifying each template is time consuming, template quantities are usually estimated based on standard purification scale and protocols to expedite the sequencing process. The sequencing method used is based on the original "Sanger" (dideoxy chain termination) sequencing methods developed in the 1970s. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001

data handling: Before the advent of capillary-based automated sequencers, DNA sequence data was produced at a comparative inefficient rate, therefore data handling and analysis were not as great a concern. DNA sequencing data production rates of a sing lab can now exceed one million bases per day. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001

Related terms: **Bioinformatics glossary**

***de novo* sequencing:** Determination of sequences (of genes or amino acids) whose sequence is not yet known. Can be done wi **LC/MS/MS** or **nanoelectrospray MS/MS.** [CHI Proteomics report]

From the Latin "*de novo*" from the beginning. See also **Mass spectrometry glossary.**

depth: See under coverage

detection methods: When DNA sequencing was first developed in the 1970s, detection methods were relatively primitive compa to today's technology. Instead of fluorescent tags, DNA fragments were labeled with radioactive tags. DNA fragments were then resolved, based on size, on vertical polyacrylamide gels and the gels exposed to X-ray film to capture the DNA fingerprint image

the separated fragments. The X-ray film was viewed by a person and scored manually. Sequencing was limited to only 200- 300 bases of data from a single sample analysis and many reiterative steps were required to acquire just a few thousand bases of DN sequence data. Development of fluorescent detection technology has allowed the data collection and analysis processes to be streamlined into a single step. DNA is separated on an automated DNA sequencer and the fluorescent image of DNA migrating through the acrylamide gel is captured by **CCD cameras** similar to that found in common home video recorders. Software specifi the automated sequencer automatically tracks and interprets the image (i.e., determines the fluorescent identity and hence the specific nucleotide at the end of each fragment), processes the data, and "calls", or interprets, the order of bases in the DNA sequence. The data can then be used in downstream sequence assembly processes. [CHI *High Throughput Genomics*] Genomi Report, Dec. 2001

dideoxy sequencing: See Sanger sequencing under Maxam-Gilbert & Sanger.

directed sequencing: See under shotgun sequencing

draft genome sequence [human]: The sequence produced by combining the information from individual sequenced clones (by creating merged sequenced contigs and them employing linking information to create scaffolds) and positioning the sequence alc the physical map of the chromosomes. (Nickname "golden path".) [Univ. of California Santa Cruz Human Genome Project Workir Draft terminology] http://genome.ucsc.edu/goldenPath/term.html

Sequence with lower accuracy than a finished sequence; some segments are missing or in the wrong order or orientation. [Histor the Human Genome Project" A Genome Glossary" Science 291: pullout chart Feb. 16, 2001]

NHGRI Rapid Data Release Policies, NHGRI, US, 2003 http://www.genome.gov/page.cfm?pageID=10506537

See also: working draft, human sequence

dynamic programming methods: Assure the optimal **global (Needleman and Wunsch** 1970; Sankoff and Kruskal 1983) or **loc** (Smith, et al. 1981) **alignment** by simply exploring all possible alignments and choosing the best. ["Pedestrian guide to analysing sequence databases" Burkhard Rost; Reinhard Schneider, 1999] http://cubic.bioc.columbia.edu/papers/1999_pedestrian/paper.hl

These methods allow the introduction of artificial gaps in aligned sequences to create an optimal alignment.

Related terms alignment, gap penalties, global alignment, local alignment, Needleman-Wunsch, sequencing algorithms

flow cytometry: Cell biology glossary

FASTA: The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scan the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculate ("init1"). Later the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable which specifies the size of a "word". (Pearson and Lipman) [NCBI Bioinformatics] More rigorous and slower than BLAS http://fasta.bioch.virginia.edu/

finished sequence - human: Sequence in which bases are identified to an accuracy of no more than 1 error in 10,000 and are placed in the right order and orientation along a chromosome with almost no gaps. [History of the Human Genome Project" A Ger Glossary" Science 291: pullout chart Feb. 16, 2001]

Each base pair has been sequenced 8-10 times, with the remaining gaps limited by present technology. ... No eukaryotic genom sequenced so far has been totally sequenced - current technology isn't up to it. Highly repetitive regions (not expected to contain many protein- coding genes) can be impossible (or very difficult) to clone. One definition of "finished" is that fewer than one ba 10,000 is incorrectly assigned. [Peer Bork, Richard Copley "Filling in the gaps" Nature 409: 218-820, 15 Feb. 2001]

At some level it's a little arbitrary when you declare a sequence essentially complete." says NHGRI Director Francis Collins... The

definition of *finished* is evolving. Our definition today is different from 10 years ago. Ten years ago we didn't even think at the lev genomes." says Laurie Goodman, editor of *Genome Research.* "I think the community at large should define *done.* Not everyone going to agree, but when you're using the word you should define what it means." Francis Collins says "You're done when you've exhausted the standard methods for closing the gaps. There should be some biological reason why those last bits of sequence e you – not because you just didn't bother." ["Are we there yet?" The Scientist :12 July 19, 1999] http://www.the-scientist.com/yr1999/july/hopkin_p12_990719.html

Related terms finished clone, Human Genome Project, post-genomic. Genomics glossary

finishing standards - Human Genome Project: The International Human Genome Consortium recognizes the need to maxi the likelihood that the finished human genome sequence meets consistent standards of quality across all participating genome centers, and to adopt uniform practices and annotation for regions that present problems for current sequencing technology. At t Seventh International Meeting, the Consortium approved a detailed set of consensus standards for what should be considered a finished sequence, a set of rules for dealing with regions that are difficult to resolve, and a set of finishing annotation tags to be submitted with accessions. Finishing Standards for the Human Genome Project - Version September 7, 2001, Standard Finishin Practices and Annotation of Problem Regions for the Human Genome Project (Genome Sequencing Center, Washington Univ. i Louis School of Medicine, US) http://www.genome.wustl.edu/Overview/finrulesname.php?G16=1

flow sorting: Cell biology glossary

full shotgun coverage: The coverage in random raw sequence needed from a large-insert clone to ensure that it is ready for finishing; this varies among centers but is typically 8-10 fold. Clones with full shotgun coverage can usually be assembled with on handful of gaps per 100 kb. [Univ. of California Santa Cruz Human Genome Project Working Draft terminology] http://genome.ucsc.edu/goldenPath/term.html

GRAIL: Gene Recognition and Assembly Internet Link software http://compbio.ornl.gov/Grail-1.3/ A suite of tools designed to pr analysis and putative annotation of DNA sequences both interactively and through the use of automated computation. [Grail Overview, Oak Ridge National Lab, US] http://compbio.ornl.gov/manuals/grail1.3-genquest.9605.shtml#GrailOverview

Does this name refer in some way to Walter Gilbert's description of the Human Genome Project as the "Holy Grail" of molecular biology?

gap: A space introduced into an **alignment** to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the ga score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the **scoring** of an alignment. [NCBI Bioinformatics]

Narrower term gap penalties

gap penalties: An important problem is the treatment of gaps, i.e., residue inserted (or deleted) to optimise the objective function Usually, gap penalties (cost of inserting and extending gaps) are chosen to be length dependent. Typically, the cost of extending gap (gap elongation) is 5-10 times lower than is the cost for introducing a gap (gap open). The optimal choice of gap penalties depends on the particular method and, in detail, on the particular sequence family ["Pedestrian guide to analysing sequence databases" Burkhard Rost, Reinhard Schneider, 1999] http://cubic.bioc.columbia.edu/papers/1999_pedestrian/paper.html

Related terms alignment, dynamic programming methods. Broader term gaps

genotype: The genetic constitution of an organism as revealed by genetic or molecular analysis, i.e. the complete set of genes, b dominant and recessive, possessed by a particular cell or organism. [IUPAC Biotech]

The observed **alleles** at a genetic **locus** for an individual. [NHLBI]

An organism's genetic makeup, as revealed through molecular analysis. [CHI SNPs Update report]

genotype to phenotype: Genomics glossary

genotyping: The process of assessing genetic variation present in an individual. CHA Cambridge Healthtech Advisors, Clinical Genomics: The Impact of Genomics on Clinical Trials and Medical Practice report, 2004

Used for diagnosis, drug efficacy, and toxicity. Utilizes genomic DNA that, after digestion, reacts with a SNP array to obtain an individual SNP pattern. These variations can for instance provide information about the diagnosis of a certain disease, or the effectiveness or side effect of a certain drug. ***Microarrays in Medicine: Arrays of Possibilities April 26-27, 2004, Boston, Massachusetts***

The determination of relevant nucleotide- base sequences in each of the two parental chromosomes. [CHI SNPs Update report]

May refer to identifying one or more, up to the entire gene sequence of an organism. Compare **phenotype**

Genotyping implies (though I haven't found this in print) determining known variants, as opposed to discovery of new ones.

Related terms Genetic variations glossary; Broader term sequencing; Narrower terms: haplotyping, viral genotyping

genotyping technologies: The ideal SNP genotyping system would offer assay uniformity, high throughput capacity, ease of as: design, and most importantly, accurate and reliable results. Various methodologies exist for detecting the alternative alleles of SN but each has its own advantages and disadvantages, and offers different levels of throughput capacity ... There are essentially thi different possibilities: 1) few SNPs, many samples, 2) many SNPs, few samples, and 3) many SNPs, many samples. While these categories may seem obvious, it is critical when optimizing a genotyping project to determine which of these categories is most important, since different technologies work best for each group. ... high- throughput [genotyping methods] analysis have only recently been optimized. Even so, there is still no "gold standard" and different companies have devised their own technologies fc genotyping. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001.

Genotyping technologies have proliferated rapidly in recent years, and at least one hundred methods are currently available for detecting the genotypes of individual SNPs. In diploid organisms, such as humans, the linkage of particular SNP genotypes on ea chromosome in a homologous pair (the haplotype) may provide additional information not available from SNP genotyping alone. [Lawrence Berkeley Lab "High Throughput Haplotying of Diploid Organisms, 2001] http://www.lbl.gov/Tech-Transfer/collaboration/techs/lbnl1748.html

"Variations on a gene: diverse technologies converge to detect human genetic variation: Amy L. Francis, Scientist 14 (15): 35, Jul 2000http://www.the-scientist.com/yr2000/jul/profile1_000724.html

glass capillary electrophoresis: Chromatography & electrophoresis glossary A type of automated **sequencer.**

global alignment: The alignment of two nucleic acid or protein sequences over their entire length. [NCBI Bioinformatics]

Related term: dynamic programming methods, Broader term alignment

Golden Path: The assembled genome sequence. Contigs are the principal building blocks of "Golden path". The term was origin: applied to the human genome assemblies coordinated at UCSC, but some people now use it for any genome assembly. Ensembl Glossary http://vega.sanger.ac.uk/Mus_musculus/helpview?se=1&kw=glossary

haplotype: A particular pattern of sequential **SNPs** found on a single chromosome. These SNPs tend to be inherited together ov time and can serve as disease-gene markers. The examination of single chromosome sets (haploid sets), as opposed to the usu chromosome pairings (diploid sets), is important because mutations in one copy of a chromosome pair can be masked by normal sequences present on the other copy. [CHI *Predictive Pharmacogenomics* report]

The linear, ordered arrangement of alleles on a chromosome. Haplotype analysis is useful in identifying recombination events. [NHLBI]

The genetic constitution of individuals with respect to one member of a pair of allelic genes, or sets of genes that are closely linke and tend to be inherited together such as those of the MAJOR HISTOCOMPATIBILITY COMPLEX. [MeSH, 1987]

The resulting **genotype** for a chromosomal **locus** constructed by combining the allele assignments from two or more genetically linked allele systems. Each individual will have two haplotypes for the **'megalocus'**, representing the **markers** on the two **homologous chromosomes**, and the haplotypes will segregate in a pedigree following Mendelian inheritance. [C. Helms, Washington Univ. St. Louis, US "Haplotyping" Nov. 1990] http://hdklab.wustl.edu/lab_manual/14/14_4.html

From "haploid genotype." The key idea is that alleles often travel in packs. This offers the hope that pharmacogenomics will not hopelessly fragment pharmaceutical fragments.

Related terms: haplotyping, haplotyping technologies Cell biology glossary diploid, haploid, ploidy; Maps: haplotype m HapMap; Narrower term: SNPs & genetic variations glossary haploinsufficiency, haplotype block, SNP haplotype

haplotyping: Somatic cells, as opposed to germ cells, have two copies of each chromosome. A given single- base position may **homozygous** for the **wild- type** base (each chromosome has the normal **allele**), homozygous for a **SNP** base (each chromosom has the altered allele), or **heterozygous** for two different bases (one chromosome has the normal allele and the other has the abnormal allele).

Haplotyping involves grouping subjects by **haplotypes**, or particular patterns of sequential SNPs, found on a single chromosom These SNPs tend to be inherited together over time and can serve as disease- gene markers. The examination of single chromos sets (**haploid** sets), as opposed to the usual chromosome pairings (**diploid** sets), is important because mutations in one copy of chromosome pair can be masked by normal sequences present on the other copy. [CHI SNPs Update report]

Genes tend to travel in packs. This is good news for **pharmacogenomics**.

Broader terms genotyping, sequencing

Haplotyping : A key approach to studying genetic variation, interview with Mark Daley, Whitehead Institute, CHI's GenomeLi 14.2 http://www.healthtech.com/newsarticles/issue14_2.asp

haplotyping technologies: Include **microarrays, mass spectrometry, sequencing**

Hidden Markov Models HMM: Molecular Modeling Useful for analyzing protein sequences.

high-throughput sequencing: For high throughput sequencing to work effectively, the DNA sequencing **templates** must be puri However, template purification has historically been a bottleneck due to its labor- intensiveness: hundreds or thousands of templa are generated for a single round of DNA sequencing, and consistency in purification methods is vital. A number of "assembly line' purification platforms have been developed that replicate the steps used in a manual template purification protocol, yet provide a greater level of consistency. Almost all of these purification platforms work flexibly with 96- and 384- well microtiter plate formats, can accommodate a wide range of purification protocols and DNA template sources. ... A high throughput lab (a high throughput academic genomics facility, or a commercial site would have between two and ten ABI 3700's or MEGABace sequencers. These would be sequencing large collection of clones or PCR products from genome screening projects and **mapping** projects. These l capacity labs would also be capable of sequencing large DNA clones such **BACs** and PACs: Finally, a super capacity sequencing such as Celera Genomics or the **Sanger Centre**, would have between 10 and 300 sequencers, with the capacity to take on full- genome sequencing projects of various **organisms**. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001.

Compare low throughput sequencing, medium throughput sequencing.

homology: Narrower terms sequence homology, sequence homology- nucleic acid; Functional genomics glossary homology Related terms homolog (homologue), similarity, ortholog, paralog, xenology; Molecular modeling homology modeling

human sequence: See draft sequence, finished sequence, published sequence, working draft

initial sequence contigs: Derived from sequenced clones [David Galas "Making sense of the sequence" Science 291: 1257-126 16 Feb. 2001]

library; library, genomic: Cell biology glossary

local alignment: The alignment of some portion of two nucleic acid or protein sequences. [NCBI Bioinformatics]

Best alignment method for sequences for whom no evolutionary relatedness is known. See **Smith- Waterman alignment.** Comp **global alignment.**

low throughput sequencing: A low throughput lab, such as a small academic lab, most often has one sequencer (perhaps a sin gel based unit that still uses radioactivity) or a core facility available (a single low capacity fluorescent sequencer). Capacity is low sequencing of few DNA clones occurs. For activities beyond that, a small academic lab would have to sub- contract to a commerc sequencing source. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001. Compare **medium throughput sequencing high throughput sequencing.**

MALDI-TOF: Mass spectrometry glossary

masking: Also known as **filtering.** The removal of repeated or low complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence. [NCBI Bioinformatics]

Maxam-Gilbert sequencing & Sanger sequencing: The two basic sequencing approaches, Maxam- Gilbert and Sanger, differ primarily in the way the nested DNA fragments are produced. Both methods work because gel electrophoresis produces very hi resolution separations of DNA molecules; even fragments that differ in size by only a single nucleotide can be resolved. Almost al steps in these sequencing methods are now automated. Maxam- Gilbert sequencing (also called the **chemical degradation meth** uses chemicals to cleave DNA at specific bases, resulting in fragments of different lengths. A refinement to the Maxam- Gilbert method known as **multiplex sequencing** enables investigators to analyze about 40 clones on a single DNA sequencing gel. Sar sequencing (also called the **chain termination** or **dideoxy method**) involves using an enzymatic procedure to synthesize DNA chains of varying length in four different reactions, stopping the DNA replication at positions occupied by one of the four bases, ar then determining the resulting fragment lengths. [Primer on Molecular Genetics, Oak Ridge National Lab, US] http://www.oml.gov/hgmis/publicat/primer/intro.html

medium throughput sequencing: A medium throughput lab (e.g., a small pharma or a university wide core facility) most often ha one to two small automated sequencers, using them to sequence a small collection of clones or PCR products for comparison sequencing. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001. Compare **low throughput sequencing, high throughput sequencing.**

megalocus: See under haplotype

merged sequence contigs: Derived by merging sequence contigs from overlapping sequenced clones. [David Galas "Making se of the sequence" Science 291: 1257-1260, 16 Feb. 2001]

microsequencing: Sequencing of proteins or peptides in very small amounts (sub microgram), sometimes for use as **probes.**

minisequencing: A **solid- phase** method for the detection of any known **point mutation** or allelic variation of DNA. In the metho amplified, **biotinylated** DNA sequences containing the mutation site are immobilized onto **streptavidin** coated microplate and pri extension reactions are carried out using labeled **nucleotides.** Incorporation of the labeled nucleotide is dependent on the **genot** and is analyzed using ELISA technique. Assay method allows automation. [Photometry applications, Labsystems Oy, Finland, no longer on website]

single base sequencing. **Related terms: single base extension array;** Genetic variations glossary

multiple sequence alignment: An alignment of three or more sequences with gaps inserted in the sequences such that residue common structural positions and/ or ancestral residues are aligned in the same column. ClustalW is one of the most widely used multiple sequence alignment programs. [NCBI Bioinformatics]

The concept of **dynamic programming** cannot be extended to align more than three sequences optimally (Murata 1990). A way around this problem is to first find optimal pairwise alignments and to then merge the pairs .["Pedestrian guide to analysing seqt databases" Burkhard Rost, Reinhard Schneider, 1999] http://cubic.bioc.columbia.edu/papers/1999_pedestrian/paper.html

Related term Hidden Markov Models HMM

multiplex sequencing: . The many applications of multiplex sequencing technology include the testing of any sample carrying **nucleic acid** material from a wide range of sources including: human clinical samples (e.g. screening for high risk subtypes of th human papilloma virus, screening blood donations for infectious organisms, identifying human **genetic variations** and **SNPs**, an selecting participants for **clinical trials**), environmental samples, the food industry, and agribusiness. ***Nucleic-Acid Based Technologies: Profiling PCR*** **June 24-26, 2002 • Washington, DC**

See also under Maxam- Gilbert sequencing.

nanopore sequencing: A nanopore can probe a long polymer of DNA and directly convert sequence information into electrical signals at very high speed. Because nanopore probing combines the power of single molecule approaches with the abilities of a high throughput system, it can be applied to SNP genotyping and multiplex haplotyping of genomic samples without the need to amplify selected molecules of interest. Using recently perfected solid- state nanopores, future applications could include rapid, complete, but inexpensive sequencing of a human genome. "Nanopore Sequencing: Implications in Personalized Medicine" Dr. D Branton, Harvard University ***Human Genome Discovery*** **Molecular Medicine Marketplace Mar. 17-19, 2003, Santa Clara CA**

Biological membrane pores are also being investigated for rapid single DNA molecule analysis. These nanometer- sized pores a constructed from a -hemolysin channels, isolated from bacteria, placed in Teflon horizontal bilayers. A single DNA molecule is pu through a nanopore in only hundreds of microseconds. The major challenge for this technology is developing the capability of rea the genetic code of the DNA fragment in the brief time that it is traversing the nanopore. [CHI *High Throughput Genomics*] report 2001.

Related term: Nanoscience & Miniaturization glossary nanopore

Needleman-Wunsch: Global sequence alignment algorithm. [Needleman, S. B., Wunsch, C. D., "A general method applicable to search for similarities in the amino acid sequence of two proteins" J. Mol. Biol.(48): 443-453 Mar. 1970] Related terms **dynamic programming; Algorithms & data management glossary, Molecular modeling glossary**

"online" automated DNA sequencers: These instruments integrate reaction thermocycling, sample purification, capillary electrophoresis, and signal detection into a single closed loop system. The DNA sample is simply injected at the beginning of the and all the protocols are automatically performed without additional human intervention. [CHI *High Throughput Genomics*] Genom Report, Dec. 2001.

optimal alignment: An alignment of two sequences with the highest possible score. [NCBI Bioinformatics]

Alignments are intended to unravel evolutionary **pathways** and/ or structural **homology** between two proteins. These two objecti (functional/ structural) may be mutually contradictory, i.e., the **'optimal' alignment'** may differ according to the objective. Yet anot perspective is the 'mathematical' optimal alignment. This is the alignment that optimises a given objective function, e.g., to find th alignment with the highest number of pairwise identical residues. **FASTA** and **BLAST** are not guaranteed to find such a mathematically optimal alignment. ["Pedestrian guide to analysing sequence databases" Burkhard Rost, Reinhard Schneider, 19 http://cubic.bioc.columbia.edu/papers/1999_pedestrian/paper.html

pathogen sequencing: In the future, more pathogens will have their genomes completely sequenced to determine not only how pathogen causes disease, but what, if any, treatments will be most effective. The DNA sequences of viruses like HIV, human

papilloma virus (HPV), and hepatitis C (HCV) are already being characterized and therapies prescribed based on this genetic information. To perform these types of diagnoses, DNA sequencing will have to become faster, more cost effective, simpler to perform, and more accessible to clinical laboratories. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001.

Phrap: Assembler software. http://www.phrap.com/background.htm

Phred: Base calling program for DNA sequence traces; ... developed by Drs. Phil Green and Brent Ewing, and is distributed und license from the University of Washington. http://www.phrap.org/

protein sequence: **Proteins glossary**

protein shotgun sequencing:

Broader term: shotgun sequencing

published working drafts - human genome: International Human Genome Sequencing Consortium special issue: **Nature** 409 (6822) 15 Feb 2001 http://www.nature.com/nature/journal/v409/n6822/ http://www.nature.com/genomics/human/papers/analysis

Human Genome [Celera Genomics sequence] special issue: **Science** 291 (5507) Feb. 16, 2001 http://www.sciencemag.org/content/vol291/issue5507/index.shtml

random sequencing: See under shotgun sequencing

resequencing: Previously sequenced site is resequenced for SNP discovery or other purposes. [CHI SNPs report]

Eric Lander, director of the Whitehead Institute's Center for Genome Research, and professor of biology at MIT notes " The huma genome will need to be sequenced only once, but it will be resequenced thousands of times, in order, for example to unravel the polygenic factors underlying human susceptibilities and predispositions ... Re-sequencing will also provide the ultimate tool for genotyping studies" [E. Lander "The New Genomics" Science 274: 536, 25 Oct. 1996]

Related terms: finished sequence, finishing standards, published working drafts, rough drafts - human genome, working drafts

SNP Single Nucleotide Polymorphism: **SNPs & Genetic Variations**

SNP discovery: Involves finding new SNPs. ... tools are just beginning to emerge and many more robust technologies are neede [NIH, Methods for Discovering and Scoring Single Nucleotide Polymorphisms, Request for Applications Jan. 9, 1998] http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-98-001.html

SNP scoring: Involves methods to determine the **genotypes** of many individuals for particular SNPs that have already been discovered. ... tools are just beginning to emerge and many more robust technologies are needed. [NIH, Methods for Discoverin and Scoring Single Nucleotide Polymorphisms, Request for Applications Jan. 9, 1998]http://grants.nih.gov/grants/guide/rfa-files/R HG-98-001.html

Sanger sequencing: See under Maxam-Gilbert sequencing.

scaffolds: Ordered set of **contigs** placed on the chromosome. [NCBI, Human Genome Home "Contig Assembly Process" Glossa Feb. 2001] http://www.ncbi.nlm.nih.gov/genome/guide/build.html#glossary.

A series of contigs that are in the right order but are not necessarily connected in one continuous stretch of sequence. [History of human Genome Project" A Genome Glossary" Science 291: pullout chart Feb. 16, 2001]

The definition of a **scaffold** appears to be quite different in the Science and Nature **draft published sequences.** [David Galas "Making sense of sequence" Science 291: 1257- Feb. 16, 2001] This is also different from the **scaffold** defined in **Drug discov and development glossary**.

The result of connecting contigs by linking information, such as paired-end reads from plasmids, paired-end reads from BACs, k mRNAs, or other sources. The contigs in a scaffold are ordered and oriented with respect to one another. [Univ. of California Sa Cruz Human Genome Project Working Draft terminology] http://genome.ucsc.edu/goldenPath/term.html

Narrower terms: sequence- contig scaffold, sequenced- clone- contig scaffold Related term: contig assembly.

scanning, scoring: SNPs & other genetic variations glossary

scoring methods: Many choices, best choice often problem dependent.. Nice review "Sequence Analysis: Which scoring metho should I use? Pittsburgh Supercomputing Center, Carnegie Mellon Univ. 1999] http://www.psc.edu/research/biomed/homologous/scoring_primer.html

Related terms filtering, gap, masking Molecular modeling glossary homology modeling Narrower term: SNP scoring

sequence alignment: The arrangement of two or more amino acid or base **sequences** from an organism or organisms in such way as to align areas of the sequences sharing common properties. The degree of relatedness or **homology** between the seque is predicted computationally or statistically based on weights assigned to the elements aligned between the sequences. This in tu can serve as a potential indicator of the genetic relatedness between the organisms. [MeSH, 1991] **Broader term?: alignments**

sequence analysis: Sequence analysis is a robust field, and mining sequence data using **bioinformatics** is one of the main acti of **genomics- based drug discovery.** Using sequence analysis to understand **whole genomes** may provide an important advan for groups looking for new drug targets among genes, or trying to pick the best among **targets** they already have. Sequence ana is one of the most widely used techniques in **genomics**. A great deal of sequence work will continue to be done, as researchers f he gaps left in the genome maps of humans and other important organisms. Studies to confirm sequence, and to identify SNPs, also need to continue. [CHI Bioinformatics report].

sequence homology: The degree of **similarity** between sequences. Studies of **amino acid** and nucleotide sequences provide useful information about the genetic relatedness of certain species. [MeSH, 1993]

Broader term Functional genomics glossary homology; Related terms Functional genomics glossary evolutionary homology; Proteomics glossary regulatory homology; Molecular modeling glossary homology modeling; Structural genomics glossary structural homology

sequence homology, amino acid The degree of similarity between sequences of amino acids. This information is useful for the understanding of genetic relatedness of certain species. [MeSH, 1993]

sequence homology - nucleic acid: The sequential correspondence of nucleotide triplets in a nucleic acid molecule which perm nucleic acid hybridization. Sequence homology is important in the study of **mechanisms** of oncogenesis and also as an indication the evolutionary relatedness of different organisms. The concept includes **viral homology.** [MeSH, 1991]

Broader term sequence homology

sequence tags: Sequence bits 2-4 contig residues in length. Used to determine the mass of a particular sequence. [CHI Proteor report] Can be used to search protein and EST databases with high specificity. [Blackstock & Weir "Proteomics" Trends in Biotechnology 17:121 Mar 1999]

sequence validation:

Sequence-contig scaffold: Scaffold produced by connecting a maximal set of sequence contigs joined by bridged gaps. [Univ. o

California Santa Cruz Human Genome Project Working Draft terminology] http://genome.ucsc.edu/goldenPath/term.html

sequenced-clone-contig scaffold: Scaffold produced by joining sequenced clone contigs by bridged SCC gaps. [Univ. of Calif Santa Cruz Human Genome Project Working Draft terminology] http://genome.ucsc.edu/goldenPath/term.html

sequencers- automation: The types and degree of automation needed at each [sequencing] step can vary greatly from laborat laboratory. The ultimate goal of automation is not to replace the role of humans, but to increase the speed and accuracy rates of individual steps while decreasing the number of repetitive manual steps that would otherwise be imposed on personnel. In fact, s of the steps may still be performed manually if that step can be performed faster or more proficiently by hand (e.g., preparation o reaction mixes, loading sequencing gels). [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001.

Related term: online automated DNA sequencers

sequencers- miniaturization: In the future, miniature DNA sequencers may become state of the art. Pocket or brief case- sized sequencers would permit onsite research to be performed in remote areas and dramatically increase the speed at which DNA fragments are analyzed. [CHI *High Throughput Genomics*] Genomic Report, Dec. 2001.

Related terms: Microarrays categories; lab- on- a- chip

sequencing: (proteins, nucleic acids) Analytical procedures for the determination of the order of **amino acids** in a polypeptide ch or of **nucleotides** in a DNA or RNA molecule. [IUPAC Compendium]

For mutation detection uses well- characterized genes to search for mutations to screen for genetic abnormalities. This process c be accelerated with the aid of oligonucleotide arrays. With this tool an individual's risk of developing certain forms of cancer or monogenetic diseases can be estimated. ***Microarrays in Medicine: Arrays of Possibilities April 26-27, 2004, Boston, Massachusetts***

Sequencing of biomolecules began with the insulin B-chain - a thirty residue peptide - which Saenger and Tuppy deduced through a combination of limited proteolysis and chemical analysis in 1951. It was a full 14 years later, until Holley et al. determined the sequence of alanine tRNA from yeast. And it took another 12 years, until "real" DNA sequencing was developed by Maxam & Gilbert and Saenger *et al* in 1977. [Introduction to bioinformatics, Univ. of Munich Gene Center, Germany, Summer 2000] http://www.lmb.uni-muenchen.de/groups/bioinformatics/01/ch_01_1.html

Largely automated now. Full DNA sequencing is the "gold standard" for **genotyping**.

Narrower terms: resequencing, sequencing - algorithms, sequencing, advanced, sequencing - cost of, sequencing - high throughput, sequencing - throughput, shotgun sequence, single DNA molecule sequencing, whole genome shotgun sequencing, chain termination sequencing, chemical cleavage sequencing, chemical degradation sequencing, *de novo* sequencing, dideoxy sequencing, microsequencing, minisequencing, multiplex sequencing, Sanger sequencing Related terms: genotyping, haplotyping

sequencing, advanced: Jay Shendure et. al., Advanced Sequencing Technologies: Methods and Goals, Nature Reviews Genet 5: 335- May 2004 http://arep.med.harvard.edu/PGP/Shendure04.pdf

sequencing algorithms: See BLAST, FASTA, Needleman - Wunsch, Smith - Waterman

sequencing - cost of: The cost of sequencing a single DNA base [when the Human Genome Project was initiated] was about $1 then; today, sequencing costs have fallen about 100-fold to $.10 to $.20 a base and still are dropping rapidly. [Human Genome N 11 (1-2) Nov. 2000] http://www.ornl.gov/hgmis/publicat/hgn/v11n1/01giants.html

sequencing - high- throughput: Uses **robotics**, automated DNA- sequencing machines and computers.

sequencing - throughput: Production of genome sequence has skyrocketed over the past year, with more than 90 percent of th sequence having been produced in the past 15 months alone. Because of this increased capacity, the next phase is expected to

much more rapidly than previously expected. [NHGRI, "International Human Genome Sequencing Consortium Publishes Sequel and Analysis of the Human Genome" Washington, D.C., February 12, 2001] http://www.nhgri.nih.gov/NEWS/Initial_sequencePR

shotgun sequencing method: Sequencing method which involves randomly sequencing tiny cloned pieces of the genome, wit foreknowledge of where on a chromosome the piece originally came from. This can be contrasted with **"directed" [sequencing]** strategies, in which pieces of DNA from adjacent stretches of a chromosome are sequenced. Directed strategies eliminate the ne for complex reassembly techniques. Because there are advantages to both strategies, researchers expect to use both random (c shotgun) and directed strategies in combination to sequence the human genome. [DOE]

Uses dynamic programming methods.

Narrower terms: chromosome-specific shotgun sequencing, protein shotgun sequencing, whole genome shotgun sequencing, YAC shotgun sequencing; Related term: full shotgun coverage

Shotgun sequencing comes of age, Tabitha Powledge, Scientist Dec. 31, 2002 http://www.biomedcentral.com/news/20021231 Hybrid of whole genome shotgun and clone- by- clone approach is probably best.

similarity: Functional genomics glossary

similarity search: BLAST, FASTA and **Smith- Waterman** are examples of similarity search **algorithms.**

single base extension array: Single Base Extension [SBE] or mini- sequencing, was one of the first genotyping technologies developed, and it is categorized as a type of primer extension method. SBE is very similar to DNA sequencing, with the exception only one base (the polymorphic base or SNP) is queried, while sequencing can identify up to several hundred bases and determi their relative order. The system is cheaper and can be performed in higher throughput than traditional sequencing. CHA Cambrid Healthtech Advisors, Clinical Genomics: The Impact of Genomics on Clinical Trials and Medical Practice report, 2004

single DNA molecule sequencing: The evolution of technology for single DNA molecule sequencing will ultimately permit whol genome analysis of populations of cells at high resolution and will obviate current PCR- based approaches, particularly important sequencing **diploid** or **polyploid cells.** This is the ultimate in sensitivity, and perhaps difficulty. Further in the future, it might be possible to utilize the protein synthesis machinery of the cell as a "sequencing engine." [National Center for Research Resource "Integrated Genomics Technologies Workshop Report" Jan 1999]

Smith-Waterman alignment: An amino acid sequence alignment that illustrates sequence similarity. The alignment is generated using the Smith- Waterman algorithm (Temple Smith and MS Waterman, J Mol Biol. 147: 195-197, 1981; WR Pearson Genomics 11:635-650, 1991) [SGD *Saccharomyces* Genome Database glossary, Stanford Univ.] http://genome-www.stanford.edu/Saccharomyces/help/glossary.htm

Related terms dynamic programming; Algorithms glossary, In silico & Molecular modeling glossary

templates: Used for sequencing generally come in two forms: 1) PCR products and 2) cloned DNA. PCR (polymerase chain reac products are derived by the PCR process where a specific but minute portion of the genome is selectively amplified 1 billion- fold the source DNA (usually a complete genome). PCR products are generated when only a small but discrete portion of the genome needs to be analyzed in hundreds or thousands of individuals. ... DNA clones are derived from cutting large portions of DNA (sometimes a complete genome) into discrete fragments that are "cloned" or inserted into DNA vectors. ... A single DNA fragment inserted into a vector and transferred into a host (generally *E. coli* bacteria) where the vector and the DNA fragment replicate as t host replicates, thus producing mass quantities of a single DNA fragment that can be purified from the host. [CHI *High Throughpu Genomics*] Genomic Report, Dec. 2001.

viral genotyping: Genomic data is enabling researchers to predict a patient's response to therapy based on the viral genotype fo viral infections. HIV genotyping is an early example of how treatment decisions are made based on the genotype of the virus.

Viral homology: See under sequence homology- nucleic acid

whole genome shotgun sequencing: Celera's whole genome shotgun sequencing technique involves sequencing from both e the double stranded cloned DNA. Celera's accurately paired clone end sequences are a key tool for assembling the genome mu more completely than single stranded sequencing methods allow at comparable levels of sequence coverage. Celera's paired er sequencing strategy, as part of the whole genome shotgun sequencing technique, has now produced sequence pairs from clone cover the human genome 11 times. The company believes that 99% of the human genome is represented in the cloned DNA. ["C Genomics completes sequencing phase of the genome from one human being" press release, Rockville, MD, April 6, 2000] http://www.pecorporation.com/press/prccorp040600.html

Broader term shotgun sequencing method.

"working draft, human genome sequence": This site contains a working draft of the human genome, which is over 90% compl Approximately half of the sequence is in a highly accurate 'finished' state. The other half is merely 'draft' quality. Some care must taken interpreting draft regions, but these are still often very useful to the working scientist. We encourage you to explore the wor draft with the genome browser, which displays the work of many annotators worldwide. [Human Genome Project Working Draft, U of California, Santa Cruz, US] http://genome.ucsc.edu/

This milestone was announced at the White House (Washington DC, US) on June 26, 2000. President Bill Clinton was joined by Francis Collins (National Human Genome Research Institute) and Craig Venter (Celera Genomics) and heads of the major US genome sequencing centers. Work continues to be done on annotating the sequence, but furthier celebration ensued with publica of two versions of the sequence in Feb. 2001.

Related terms: draft sequence, finished sequence - human, published working drafts; Genomics glossary Human Genor Project

YAC shotgun sequencing: Yeast artificial chromosome clones are isolated from a YAC library using chromosome- specific prob (or markers). These YACs are then used to prepare small- insert shotgun libraries. Clones from these libraries are then submit to sequencing. *Plasmodium falciparum* Genome Project Sequencing Strategies, NCBI, US, 2001 http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/pfalciparum1.html

Bibliography

CHI *Predictive Pharmacogenomics* report

DNA Sequencing glossary, Geospiza, Inc., 2002, 26 definitions. http://www.geospiza.com/community/support/glossary/

Guide to Molecular Sequence Analysis Glossary, Andrew Louka, 2001, 35 definitions http://www.brunel.ac.uk/depts/bl/project/biocomp/sequence/seqanal_guide/glossary.html

NCBI (US) BLAST Glossary, 2000. 40+ definitions http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html

Human Genome Project Information, Facts about Genome Sequencing, Oak Ridge National Lab, US, 2002 http://www.ornl.gov/hgmis/faq/seqfacts.html

Alpha glossary index

How to look for other unfamiliar terms

)

)